

# 周报

本周先是对实现完毕的基于相似哈希信息指纹的压缩程序进行真实数据效果测试。结果发现效果不好，对于其原因的分析发现是由于基于字符串相似哈希汉明距离的相似度度量不能真实的反应体数据块之间的相似程度，总的趋势是相似程度在 0.5 上下徘徊，再追根溯源，发现可能的内在原因有两个，一是当前的实现方法是先将体数据值排成一个 **uchar** 串，然后转化成 **string** 字符串，再分词，然后生成对应的 **md5** 哈希，进一步生成相似哈希信息指纹。上述先转化为 **string** 字符串的步骤造成了结果的大幅度异常。二是相似哈希这个方法对于小数据的应用确实是有天然的缺陷。目前正在寻找 **md5** 的替换方法，并尝试直接使用体数据值构成的串进行哈希生成，排除干扰。如果这种排除第一种原因的可能性后，那么相似哈希也许并不能用在我们的当前方案里了。

如果无法直接应用相似哈希，那么只能尝试使用 **knn** 图构建体数据块的邻接矩阵，因为直接使用全局联通相似性图构建邻接矩阵会使得大数据压缩时，内存开辟出现无法控制的大小限制。另外，使用体数据值向量的高斯函数距离

$$distance = e^{-(x1-x2)^2/2(c)^2}$$

构建相似性值，亦可尝试使用欧式距离。直接使用谱聚类构建码表和索引体纹理。

建刚的程序也已经完成测试，浮点数据的有损压缩压缩比达到 **80:1**，压缩质量（图像保真度）也很不错，在浮点型的 **bucky ball32\*32\*32** 数据上均方差 **mse** 为 **25** 左右，**512\*512\*512** 的原子裂变数据上 **mse** 为 **18** 左右，信噪比 **SNR** 也都在 **20** 以上。不过，残差率都较差，测试了 **4** 个数据，最差的为 **0.02**，最好的为 **0.4** 左右，所以目前无法使用 **PSH** 来压缩残差体纹理以进行近似无损压缩的构想。应用在气象数据里面已经没有问题。

对于建刚的程序，我调研了相关文献后，有三个改进方法。如下：

**1.**对于拉普拉斯分解得到的两个下采样体纹理，先分别进行向量量化，然后分别计算残差；而不是原先的全部计算完毕以后，再计算残差。这样，由于拉普拉斯分解后，实际上得到的是原始体数据的滤波后效果，分别计算的残差可能会有较大的质量改进。

这种方法已经尝试，结果是，分别提升了 **10%** 的残差率，虽然提升较大，还是不足以使用 **PSH**，要使用 **PSH**，由压缩比计算公式可得，必须使得残差率在 **0.8** 以上。

**2.**原先方法为：先使用 **PCA** 得到向量量化初始码表(**256** 个码元)，再使用 **LBGVQ** 对结果做

***refinement***，以减少残差，最后再做索引体纹理和码表，完成压缩段。目前的改进思想是：先用 ***PCA*** 得到比较多的类别分类（***1024***），这样得到的分类具有初步的分类效果，然后在此基础上，将这 ***1024*** 个码元作为新的输入，使用谱聚类方法构建 ***KNN*** 相似矩阵，实际上即以 ***1024*** 个具有初步分类效果的码元向量作为原始体数据特征点，降低体数据的等价体素点个数，这样即可避免谱聚类无法处理大数据的问题，同时，由于 ***PCA*** 的分类结果，使得特征点的选取不是盲目无效的。在谱聚类做完以后，即可得到码表和索引体纹理。这同样是一种新方法，只是拉普拉斯分解是借鉴 ***03*** 年 ***VIS*** 论文的方法，目前正在讨论算法中。

3. 直接扩大码表至 ***65536***，这种方法是最后的选择，看能否得到稀疏的残差。在之前的纯向量量化代码上尝试过，一些数据上会有比较好的结果。

教材的工作下周和陈老师讨论后开展。